Thank you Tom for your thorough and constructive review on our paper. We answer to your comments below:

For a comparison to be fair, I would recommend using caret package or tuneRF package to get a better estimate of the random forest parameters, especially "mtry". "mtry" has shown to lead to very different results but can be relatively inexpensively optimized.

We agree that the mtry parameters of the Random Forest model can be optimized and this can lead to better prediction accuracy. We will do so in the revised manuscript. Please note that mtry has already been manually tuned by trying several values. We will use an search algorithm to find maybe more appropriate values.

Points shown in Figure 2 show significant spatial clustering. Again, to get a realistic estimate of mapping accuracy I would recommend using spatial CV (see https://geocompr.robinlovelace.net/spatial-cv.html#intro-cv and https://mlr.mlrorg. com/articles/tutorial/handling_of_spatial_data.html). If it might help you to run spatial CV, we have developed a function for spatially subsetting points per fixed block (http://gsif.r-forge.r-project.org/sample.grid.html).

The sampling locations on this area are not really clustered even though Fig. 2 seems to show spatial clustering. The samples collected are based on a random catena design. Repeating the calibration based on different split between calibration and validation would simply lead to more stable results, which can be useful in some cases, e.g., for a sensitivity analysis. In our case study we do not perform such analysis. In addition, computing time for calibrating a single model (mentioned in the discussion) does not allow repeating this procedure several dozens of time. This is a limitation that we recognize in the Discussion and that must be addressed by further research. The calibration and validation sets are used to validate both random forest and CNN so that no method is disadvantaged using this split. A random split of the calibration and validation sets are very common in the soil mapping literature.

Any CV to be stable should be repeated e.g. 5-10 times. Single split could by accident lead to over-optimistic/pessimistic results.

See answer to comment 2.

Other comments:

Otherwise, I would highly recommending rewriting methods section, primarily to add more detailed explanation to non-CNN experts. I have read the paper two times and I have to admit that it was still not totally clear to me whether you fit models for top and sub-soil separately or at once. Figure 1 could probably be split into 2 figures and then you can explain some concepts of CNN somewhat clearer. You use many abstract terms from ML and these probably need a gentle intro for soil scientists.

The Method section will be modified in view on your comments and other reviewers comments. Many useful suggestions for improvements were made and we will account for them while revising the manuscript. In particular, technical terms will be clarified to non-expert readers.

Your Paper is, in essence, twin-paper of Behrens et al. (2018) and leads to similar results/conclusions (Behrens use the term "Gaussian scale space" to characterize scaling), so please emphasize more what is really different in your approach as compared to their approach. For example, Behrens et al. (2018) run more complex terrain analysis for a range of aggregated DEMs, but they do not deal with NIR/MIR data specifically etc.

We will put emphasise on the differences with the paper of Behrens (2018). Basically Behrens et al.'s deep learning is a multi-layer feedforward artificial neural network, while our model is a convolutional neural network. In addition, Behrens requires a pre-processing step: creation of multiscale data as inputs to the ANN, while CNN just accepts images as inputs and performs the multiscale analysis directly from the input.

Comments on the manuscript:

Most of them are useful small text editing comments. Some comments are similar to those made by other reviewers.

a) Line 9 page 9: We used a univariate RF model for each soil depth, i.e. a single model is trained by depth, while CNN predict both depths using a single calibrated model. We will make it clearer in the text.
b) Fig. 3: We would like to keep this figure. We believe it is interesting to visualize the covariates for interpretation of the predicted maps.
c) Line 15 page 12: for CNN we built one single model for both depths.
d) Line 5 page 14: Thank you for the suggestion. We will modify the text accordingly.
e) Line 22 page 16: Maps of the prediction error could be obtained, but this is not straightforward and difficult to implement while accounting for the measurement error. Deriving such maps would certainly need an important research effort.
f) Line 6 page 17: We agree that this is well known. The first paragraph in the discussion already mentioned this clearly and the relevant literature.